

# Generalized Wald-type Tests based on Minimum Density Power Divergence Estimators

Basu, A.<sup>1</sup>; Mandal, A.<sup>1</sup>; Martin, N.<sup>2</sup> and Pardo, L.<sup>3</sup>

<sup>1</sup>Indian Statistical Institute, Kolkata 700108, India

<sup>2</sup>Department of Statistics, Carlos III University of Madrid, 28038 Getafe (Madrid), Spain

<sup>3</sup>Department of Statistics and O.R. Complutense University of Madrid, 28040 Madrid, Spain

April 1, 2014

## Abstract

In testing of hypothesis the robustness of the tests is an important concern. Generally, the maximum likelihood based tests are most efficient under standard regularity conditions, but they are highly non-robust even under small deviations from the assumed conditions. In this paper we have proposed generalized Wald-type tests based on minimum density power divergence estimators for parametric hypotheses. This method avoids the use of nonparametric density estimation and the bandwidth selection. The trade-off between efficiency and robustness is controlled by a tuning parameter  $\beta$ . The asymptotic distributions of the test statistics are chi-square with appropriate degrees of freedom. The performance of the proposed tests are explored through simulations and real data analysis.

**AMS 2010 subject classification:** Primary 62F35; Secondary 62F03.

**Keywords:** Density Power Divergence, Robustness, Tests of Hypotheses.

## 1 Introduction

Basu et al. (1998) have introduced the minimum divergence estimator (MDPDE) that minimizes the density power divergence measure. The robustness properties of these estimators have been studied in detail by several authors. However, the problem of hypothesis testing based on the density power divergence measures has only recently been explored (Basu et al., 2013, 2014). The results indicate that the test statistics based on the density power divergence measures, when the parameters are estimated by the MDPDEs, have substantially superior performance compared to the likelihood ratio test in the presence of outliers. On the other hand, in pure data these tests are often competitive to the likelihood ratio tests. So the tests based on the density power divergence are very useful practical tools in robust statistics.

In some situations it is not easy to get the expression of the density power divergence measure between the population densities estimated under the null hypothesis and under the unrestricted parameter space. Another difficulty is that in many situations the asymptotic distribution of the test statistic is a linear combination of independent chi-square variables. While none of these problems are insurmountable, they lead to some reduction in the appeal of these otherwise useful tests.

To overcome these problems we will present a class of generalized Wald-type test statistics based on minimum density power divergence estimators of the model parameters. Our aim is to manipulate the statistics in a manner that allows us to exploit the nice properties of the MDPDEs in constructing the test statistics, and yet come up with asymptotic distributions which are pure chi-squares rather than linear combinations of independent chi-squares. On the whole, we expect to make the tests theoretically and operationally far more simple, without compromising the efficiency and the robustness properties of the tests.

In Section 2 we have introduced the necessary notations and described the existing results. We have introduced the Wald-type test statistic in the context of the simple null hypothesis in Section 3.

The problem of the composite null hypothesis is considered in Section 4. Some simulation studies are reported in Section 5 to explore the behavior of the proposed families of test statistics. Section 6 has some concluding remarks, and the proofs are given in the Appendix.

## 2 Some notational and previous concepts

Let  $\mathcal{G}$  denote the set of all distributions having densities with respect to a dominating measure (generally the Lebesgue measure or the counting measure). Given any two densities  $g$  and  $f$  in  $\mathcal{G}$ , the density power divergence between them is defined, as the function of a nonnegative tuning parameter  $\beta$ , as

$$d_\beta(g, f) = \begin{cases} \int \left\{ f^{1+\beta}(x) - \left(1 + \frac{1}{\beta}\right) f^\beta(x)g(x) + \frac{1}{\beta} g^{1+\beta}(x) \right\} dx, & \text{for } \beta > 0, \\ \int g(x) \log \left( \frac{g(x)}{f(x)} \right) dx, & \text{for } \beta = 0. \end{cases} \quad (1)$$

The case corresponding to  $\beta = 0$  may be derived from the general case by taking the continuous limit as  $\beta \rightarrow 0$ , and in this case  $d_0(g, f)$  is the classical Kullback-Leibler divergence. The quantities defined in equation (1) are genuine divergences in the sense  $d_\beta(g, f) \geq 0$  for all  $g, f \in \mathcal{G}$  and all  $\beta \geq 0$ , and  $d_\beta(g, f)$  is equal to zero if and only if the densities  $g$  and  $f$  are identically equal. More details about inference based on divergence measures can be found in Pardo (2006) and Basu et al. (2011).

We consider a parametric model of densities  $\{f_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$  and suppose that we are interested in the estimation of  $\theta$ . Let  $G$  represent the distribution function corresponding to the density  $g$ . The minimum density power divergence functional  $T_\beta(G)$  at  $G$  is defined by the requirement  $d_\beta(g, f_{T_\beta(G)}) = \min_{\theta \in \Theta} d_\beta(g, f_\theta)$ . Clearly the term  $\int g^{1+\beta}(x)dx$  in (1) has no role in the minimization of  $d_\beta(g, f_\theta)$  over  $\theta \in \Theta$ . Thus the essential objective function to be minimized in the computation of the minimum density power divergence functional  $T_\beta(G)$  reduces to

$$\int f_\theta^{1+\beta}(x)dx - \left(1 + \frac{1}{\beta}\right) \int f_\theta^\beta(x)g(x)dx = \int f_\theta^{1+\beta}(x)dx - \left(1 + \frac{1}{\beta}\right) \int f_\theta^\beta(x)dG(x). \quad (2)$$

Notice that in the above objective function the density  $g$  appears only as a linear term (unlike, say, the objective function of the minimum Hellinger distance functional where the square root of the density  $g$  is the relevant quantity). Given a random sample  $X_1, \dots, X_n$  from the distribution  $G$ , we can therefore approximate the objective function in (2) by replacing  $G$  with its empirical estimate  $G_n$ . For a given tuning parameter  $\beta$ , the MDPDE  $\hat{\theta}_\beta$  of  $\theta$  can be obtained by minimizing

$$\begin{aligned} & \int f_\theta^{1+\beta}(x)dx - \left(1 + \frac{1}{\beta}\right) \int f_\theta^\beta(x)dG_n(x) \\ &= \int f_\theta^{1+\beta}(x)dx - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^n f_\theta^\beta(X_i) = \frac{1}{n} \sum_{i=1}^n V_\theta(X_i) \end{aligned} \quad (3)$$

over  $\theta \in \Theta$ , where  $V_\theta(x) = \int f_\theta^{1+\beta}(y)dy - \left(1 + \frac{1}{\beta}\right) f_\theta^\beta(x)$ . In the special case  $\beta = 0$ , we must minimize the expression  $-\frac{1}{n} \sum_{i=1}^n \log f_\theta(X_i)$ ; the corresponding minimizer turns out to be the maximum likelihood estimator (MLE) of  $\theta$ . The minimization of the expression in (3) over  $\theta$  does not require the use of a nonparametric density estimate of the true unknown distribution  $G$ . Existing theory (e.g. De Angelis and Young, 1992) shows that in general there is little or no advantage in introducing smoothing for such functionals which may be empirically estimated using the empirical distribution function alone, except in very special cases. Using  $G_n$  to substitute  $G$ , if possible, is therefore a natural step.

Let  $\mathbf{u}_\theta(x) = \frac{\partial}{\partial \theta} \log f_\theta(x)$  be the score function of the model. Under differentiability of the model the minimization of the objective function in equation (3) leads to an estimating equation of the form

$$\frac{1}{n} \sum_{i=1}^n \mathbf{u}_\theta(X_i) f_\theta^\beta(X_i) - \int \mathbf{u}_\theta(x) f_\theta^{1+\beta}(x)dx = \mathbf{0}_p, \quad (4)$$

which is an unbiased estimating equation under the model. Since the corresponding estimating equation weights the score  $\mathbf{u}_\theta(X_i)$  with the power of the density  $f_\theta^\beta(X_i)$ , the outlier resistant behavior of the estimator is intuitively apparent. See [Basu et al. \(1998\)](#) and [Jones et al. \(2001\)](#) for more details.

The functional  $T_\beta(\cdot)$  is Fisher consistent; it takes the value  $\theta_0$  when the true density  $g = f_{\theta_0}$  is in the model. When it is not,  $\theta_\beta^g = T_\beta(G)$  represents the best fitting parameter. For brevity we will suppress the  $g$  superscript in the notation for  $\theta_\beta^g$ ;  $f_{\theta_\beta}$  is the model element closest to the density  $g$  in the minimum density power divergence sense corresponding to the tuning parameter  $\beta$ .

Let  $g$  be the true data generating density, and  $\theta_\beta = T_\beta(G)$  be the best fitting parameter. To set up the notation we define the quantities

$$\begin{aligned} \mathbf{J}_\beta(\theta) &= \int \mathbf{u}_\theta(x) \mathbf{u}_\theta^T(x) f_\theta^{1+\beta}(x) dx \\ &\quad + \int \{\mathbf{I}_\theta(x) - \beta \mathbf{u}_\theta(x) \mathbf{u}_\theta^T(x)\} \{g(x) - f_\theta(x)\} f_\theta^\beta(x) dx, \end{aligned} \quad (5)$$

$$\mathbf{K}_\beta(\theta) = \int \mathbf{u}_\theta(x) \mathbf{u}_\theta^T(x) f_\theta^{2\beta}(x) g(x) dx - \boldsymbol{\xi}_\beta(\theta) \boldsymbol{\xi}_\beta^T(\theta), \quad (6)$$

where  $\boldsymbol{\xi}_\beta(\theta) = \int \mathbf{u}_\theta(x) f_\theta^\beta(x) g(x) dx$ , and  $\mathbf{I}_\theta(x) = -\frac{\partial}{\partial \theta} \mathbf{u}_\theta(x)$  is the so called Fisher information matrix at the model. The following results, proved in [Basu et al. \(2011\)](#), form the basis of our subsequent developments. We assume that the conditions D1–D5 of [Basu et al. \(2011, p. 304\)](#) are true. Then the following results hold.

1. The minimum density power divergence estimating equation (4) has a consistent sequence of roots  $\hat{\theta}_\beta$ , i.e.  $\hat{\theta}_\beta \xrightarrow[n \rightarrow \infty]{\mathcal{P}} \theta_\beta$ .
2.  $n^{1/2}(\hat{\theta}_\beta - \theta_\beta)$  has an asymptotic multivariate normal distribution with mean zero (vector) and covariance matrix  $\mathbf{J}^{-1} \mathbf{K} \mathbf{J}^{-1}$ , where  $\mathbf{J} = \mathbf{J}_\beta(\theta_\beta)$ ,  $\mathbf{K} = \mathbf{K}_\beta(\theta_\beta)$  are as in (5) and (6).

### 3 Wald-type Statistics for the Simple Null Hypothesis

Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a distribution modeled by probability density function  $f_\theta(x)$ , where  $\theta \in \Theta \subset \mathbb{R}^p$  and  $x \in \mathcal{X}$ , the sample space. In this section we will define a family of Wald-type test statistics based on MDPDEs for testing the null hypothesis

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \neq \theta_0. \quad (7)$$

**Definition 1** Let  $\hat{\theta}_\beta$  be the MDPDE of  $\theta$ . The family of Wald-type test statistics for testing the null hypothesis in (7) is given by

$$W_n = n \left( \hat{\theta}_\beta - \theta_0 \right)^T \left( \mathbf{J}_\beta^{-1}(\theta_0) \mathbf{K}_\beta(\theta_0) \mathbf{J}_\beta^{-1}(\theta_0) \right)^{-1} \left( \hat{\theta}_\beta - \theta_0 \right), \quad (8)$$

where the matrices  $\mathbf{J}_\beta(\theta_0)$  and  $\mathbf{K}_\beta(\theta_0)$  are as defined in (5) and (6) respectively.

When  $\beta = 0$ ,  $\hat{\theta}_\beta$  coincides with the maximum likelihood estimator  $\hat{\theta}$  of  $\theta$ , and  $\mathbf{J}_\beta^{-1}(\theta_0) \mathbf{K}_\beta(\theta_0) \mathbf{J}_\beta^{-1}(\theta_0)$  coincides with the inverse of the Fisher information matrix, and thus we recover the classical Wald statistic or the ordinary Wald statistic, terms that we will use interchangeably.

The asymptotic distribution of  $W_n$  is presented in the next theorem. The result follows easily from the asymptotic distribution of MDPDE considered in Section 2, so we skip the proof.

**Theorem 2** The asymptotic null distribution of the Wald-type test statistic given in (8) is a chi-square distribution with  $p$  degrees of freedom.

In many cases the power function of this testing procedure can not be calculated explicitly. In the following theorem we present a useful asymptotic result for approximating the power function of the Wald-type test statistics given in (8).

**Theorem 3** Let  $\theta^*$  be the true value of parameter with  $\theta^* \neq \theta_0$ . Then the convergence

$$n^{1/2} \left( l(\hat{\theta}_\beta) - l(\theta^*) \right) \xrightarrow[n_2 \rightarrow \infty]{L} N(\mathbf{0}, \sigma_W^2(\theta^*))$$

holds, where

$$l(\theta) = (\theta - \theta_0)^T \left( \mathbf{J}_\beta^{-1}(\theta_0) \mathbf{K}_\beta(\theta_0) \mathbf{J}_\beta^{-1}(\theta_0) \right)^{-1} (\theta - \theta_0),$$

and

$$\sigma_W^2(\theta^*) = 4(\theta^* - \theta_0)^T \left( \mathbf{J}_\beta^{-1}(\theta^*) \mathbf{K}_\beta(\theta^*) \mathbf{J}_\beta^{-1}(\theta^*) \right)^{-1} (\theta^* - \theta_0). \quad (9)$$

**Proof.** See the Appendix. ■

**Remark 4** Now we can approximate the power,  $\beta_{W_n}$ , of the Wald-type test statistics at  $\theta^*$ , by

$$\begin{aligned} \beta_{W_n}(\theta^*) &= P_{\theta^*}(W_n > \chi_{p,\alpha}^2) \\ &= P_{\theta^*} \left( l(\hat{\theta}_\beta) - l(\theta^*) > \frac{\chi_{p,\alpha}^2}{n} - l(\theta^*) \right) \\ &= P_{\theta^*} \left( n^{1/2} \left( l(\hat{\theta}_\beta) - l(\theta^*) \right) > n^{1/2} \left( \frac{\chi_{p,\alpha}^2}{n} - l(\theta^*) \right) \right) \\ &= P_{\theta^*} \left( n^{1/2} \frac{l(\hat{\theta}_\beta) - l(\theta^*)}{\sigma_W(\theta^*)} > \frac{n^{1/2}}{\sigma_W(\theta^*)} \left( \frac{\chi_{p,\alpha}^2}{n} - l(\theta^*) \right) \right) \\ &= 1 - \Phi_n \left( \frac{n^{1/2}}{\sigma_W(\theta^*)} \left( \frac{\chi_{p,\alpha}^2}{n} - l(\theta^*) \right) \right), \end{aligned} \quad (10)$$

where  $\Phi_n(\cdot)$  is a sequence of distribution functions tending uniformly to the standard normal distribution function  $\Phi(\cdot)$ . It is clear that

$$\lim_{n \rightarrow \infty} \beta_{W_n}(\theta^*) = 1,$$

for all  $\alpha \in (0, 1)$ . Therefore, the test is consistent in the sense of Fraser ([Fraser, 1957](#)).

Based on this result given in (10) we can calculate the required sample size in order to get a given value for the power. If we want to get a power of  $\beta^* = \beta_{W_n}(\theta^*)$ , we need a sample size  $n$  given by

$$n = [n^*] + 1, \quad n^* = \frac{A + B + \sqrt{A(A + 2B)}}{2l^2(\theta^*)}, \quad (11)$$

where  $[x]$  is the largest integer less than or equal to  $x$ ,  $A = \sigma_W^2(\theta^*) (\Phi^{-1}(1 - \beta^*))^2$  and  $B = \frac{1}{2} \chi_{p,\alpha}^2 l(\theta^*)$ . Here  $\chi_{p,\alpha}^2$  is the 100(1 -  $\alpha$ ) percentile point of chi-square distribution with  $p$  degrees of freedom.

Now we are going to get the asymptotic distribution of  $W_n$  under contiguous alternative hypotheses described by

$$H_{1,n} : \theta_n = \theta_0 + n^{-1/2} \mathbf{d}, \quad (12)$$

where  $\mathbf{d}$  is a fixed vector in  $\mathbb{R}^p$  such that  $\theta_n \in \Theta \subset \mathbb{R}^p$  for all  $n$ .

**Theorem 5** Under the contiguous alternative hypotheses given in (12) the asymptotic distribution of the Wald-type test statistic  $W_n$  is a non-central chi-square with  $p$  degrees of freedom and non-centrality parameter

$$\delta = \mathbf{d}^T \mathbf{J}_\beta^{-1}(\theta_0) \mathbf{K}_\beta(\theta_0) \mathbf{J}_\beta^{-1}(\theta_0) \mathbf{d}. \quad (13)$$

**Proof.** See the Appendix. ■

**Remark 6** Using the last theorem we can get an approximation to the power function at the contiguous alternative  $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2}\mathbf{d}$  through the relation

$$\beta_{W_n}(\boldsymbol{\theta}_n) = 1 - G_{\chi_p^2(\delta)}(\chi_{p,\alpha}^2),$$

where  $G_{\chi_p^2(\delta)}(\chi_{p,\alpha}^2)$  is the distribution function of a non-central chi-square, with  $p$  degrees of freedom and non-centrality parameter  $\delta$  given in equation (13), evaluated at the point  $\chi_{p,\alpha}^2$ .

We can observe that this expression permits us to obtain an approximation of the power function in a generic point  $\boldsymbol{\theta}^*$ , because we can consider  $\mathbf{d} = n^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)$  and then  $\boldsymbol{\theta}_n = \boldsymbol{\theta}^*$ .

**Example 7** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from  $\text{Exp}(\theta)$ , the exponential distribution with mean  $\theta$ . Suppose we want to test the hypothesis  $H_0 : \theta = \theta_0$ , against  $H_1 : \theta \neq \theta_0$ . It is easy to show that the minimum density power divergence estimator  $\hat{\theta}_\beta$  of  $\theta$  can be obtained by iteratively solving the equation

$$\hat{\theta}_\beta = \frac{\sum_{i=1}^n X_i \exp\left\{-\frac{\beta X_i}{\hat{\theta}_\beta}\right\}}{\sum_{i=1}^n \exp\left\{-\frac{\beta X_i}{\hat{\theta}_\beta}\right\} - \frac{n\beta}{(1+\beta)^2}}. \quad (14)$$

Note that, putting  $\beta = 0$  in equation (14), we get an explicit expression for the MLE as  $\hat{\theta}_0 = n^{-1} \sum_{i=1}^n X_i$ . The standard calculations of the matrices  $\mathbf{J}_\beta(\boldsymbol{\theta}_0)$  and  $\mathbf{K}_\beta(\boldsymbol{\theta}_0)$ , as defined in (5) and (6) respectively, gives us

$$n^{1/2}(\hat{\theta}_\beta - \theta_0) \xrightarrow[n \rightarrow \infty]{L} N(0, h(\beta)\theta_0^2),$$

where

$$h(\beta) = \frac{(1+\beta)^2 P(\beta)}{(1+\beta^2)^2(1+2\beta)^3}, \text{ and } P(\beta) = 1 + 4\beta + 9\beta^2 + 14\beta^3 + 13\beta^4 + 8\beta^5 + 4\beta^6. \quad (15)$$

Therefore, for testing hypothesis  $H_0 : \theta = \theta_0$ , against  $H_1 : \theta \neq \theta_0$ , the Wald-type test statistic is given by

$$W_n = \frac{n(\hat{\theta}_\beta - \theta_0)^2}{h(\beta)\theta_0^2}, \quad (16)$$

whose asymptotic distribution, under the null hypothesis, is a chi-square with one degree of freedom.

## 4 Wald-type Test Statistics for the Composite Null Hypothesis

We will now consider the problem of testing composite null hypothesis. In testing of a composite null hypothesis, the restricted parameter space  $\Theta_0 \subset \Theta$  is often defined by a set of  $r$  restrictions of the form

$$\mathbf{m}(\boldsymbol{\theta}) = \mathbf{0}_r, \quad (17)$$

where  $\mathbf{m} : \mathbb{R}^p \rightarrow \mathbb{R}^r$ , and  $\mathbf{0}_r$  denotes the null vector of dimension  $r$  (see Serfling, 1980). Assume that the  $p \times r$  matrix

$$\mathbf{M}(\boldsymbol{\theta}) = \frac{\partial \mathbf{m}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (18)$$

exists and is continuous in  $\boldsymbol{\theta}$ , and  $\text{Rank}(\mathbf{M}(\boldsymbol{\theta})) = r$ , where  $r \leq p$ . Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a distribution modeled by probability density function  $f_\theta(x)$ , where  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$  and  $x \in \mathcal{X}$ , the sample space. Our interest is in testing the hypothesis

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ against } H_1 : \boldsymbol{\theta} \notin \Theta_0, \quad (19)$$

where  $\Theta_0$  is a subset of the parameter space  $\Theta$ .

**Definition 8** Let  $\hat{\theta}_\beta$  be the MDPDE of  $\theta$ . The family of Wald-type test statistics for testing the null hypothesis in (19) is given by

$$W_n = n\mathbf{m}^T(\hat{\theta}_\beta) \left[ \mathbf{M}^T(\hat{\theta}_\beta) \mathbf{J}_\beta^{-1}(\hat{\theta}_\beta) \mathbf{K}_\beta(\hat{\theta}_\beta) \mathbf{J}_\beta^{-1}(\hat{\theta}_\beta) \mathbf{M}(\hat{\theta}_\beta) \right]^{-1} \mathbf{m}(\hat{\theta}_\beta), \quad (20)$$

where the matrices  $\mathbf{M}$ ,  $\mathbf{J}_\beta$  and  $\mathbf{K}_\beta$  were defined in equations (18), (5) and (6) respectively, and the function  $\mathbf{m}$  is defined in (17).

In the special case when  $\beta = 0$ ,  $\hat{\theta}_\beta$  coincides with the maximum likelihood estimator  $\hat{\theta}$  of  $\theta$ , and  $\mathbf{J}_\beta^{-1}(\hat{\theta}_\beta) \mathbf{K}_\beta(\hat{\theta}_\beta) \mathbf{J}_\beta^{-1}(\hat{\theta}_\beta)$  which is the inverse of the Fisher information matrix. Thus the test statistic in (20) reduces to the classical Wald test statistic. In the next theorem we present the asymptotic distribution of  $W_n$ .

**Theorem 9** The asymptotic distribution of the Wald-type test statistics given in (20) is chi-square with  $r$  degrees of freedom.

**Proof.** See the Appendix. ■

Consider the null hypothesis  $H_0 : \theta \in \Theta_0 \subset \Theta$ . By Theorem 9, the null hypothesis is rejected if  $W_n \geq \chi_{r,\alpha}^2$ . The following theorem can be used to approximate the power function. Assume that  $\theta^* \notin \Theta_0$  is the true value of the parameter so that the unrestricted estimator  $\hat{\theta}_\beta \xrightarrow[n \rightarrow \infty]{p} \theta^*$ .

**Theorem 10** Suppose

$$l^*(\theta_1, \theta_2) = n\mathbf{m}^T(\theta_1) \left[ \mathbf{M}^T(\theta_2) \mathbf{J}_\beta^{-1}(\theta_2) \mathbf{K}_\beta(\theta_2) \mathbf{J}_\beta^{-1}(\theta_2) \mathbf{M}(\theta_2) \right]^{-1} \mathbf{m}(\theta_1).$$

Then

$$n^{1/2} \left( l^*(\hat{\theta}_\beta, \hat{\theta}_\beta) - l^*(\theta^*, \theta^*) \right) \xrightarrow[n \rightarrow \infty]{L} N(0, \sigma_W^2(\theta^*)),$$

where

$$\sigma_W^2(\theta^*) = \left( \frac{\partial l^*(\theta, \theta^*)}{\partial \theta} \right)_{\theta=\theta^*}^T \mathbf{J}_\beta^{-1}(\theta^*) \mathbf{K}_\beta(\theta^*) \mathbf{J}_\beta^{-1}(\theta^*) \left( \frac{\partial l^*(\theta, \theta^*)}{\partial \theta} \right)_{\theta=\theta^*}^T. \quad (21)$$

**Proof.** See the Appendix. ■

We may also find an approximation of the power of  $W_n$  at an alternative close to the null hypothesis. Let  $\theta_n \in \Theta - \Theta_0$  be a given alternative, and let  $\theta_0$  be the element in  $\Theta_0$  closest to  $\theta_n$  in terms of the Euclidean distance. One possibility to introduce contiguous alternative hypotheses in this context is to consider a fixed  $\mathbf{d} \in \mathbb{R}^p$  and to permit  $\theta_n$  move towards  $\theta_0$  as  $n$  increases through the relation

$$H_{1,n} : \theta_n = \theta_0 + n^{-1/2} \mathbf{d}. \quad (22)$$

A second approach is to relax the condition  $\mathbf{m}(\theta) = \mathbf{0}_r$  defining  $\Theta_0$ . Let  $\delta \in \mathbb{R}^r$  and consider the following sequence of parameters  $\{\theta_n\}$  moving towards  $\theta_0$  according to the set up

$$H_{1,n}^* : \mathbf{m}(\theta_n) = n^{-1/2} \delta. \quad (23)$$

Note that a Taylor series expansion of  $\mathbf{m}(\theta_n)$  around  $\theta_0$  yields

$$\mathbf{m}(\theta_n) = \mathbf{m}(\theta_0) + \mathbf{M}^T(\theta_0) (\theta_n - \theta_0) + o(\|\theta_n - \theta_0\|). \quad (24)$$

By substituting  $\theta_n = \theta_0 + n^{-1/2} \mathbf{d}$  in (24) and taking into account that  $\mathbf{m}(\theta_0) = \mathbf{0}_r$ , we get

$$\mathbf{m}(\theta_n) = n^{-1/2} \mathbf{M}^T(\theta_0) \mathbf{d} + o(\|\theta_n - \theta_0\|). \quad (25)$$

So the equivalence relationship between the hypotheses  $H_{1,n}$  and  $H_{1,n}^*$  is

$$\delta = \mathbf{M}^T(\theta_0) \mathbf{d} \text{ as } n \rightarrow \infty. \quad (26)$$

In the following theorem we show the asymptotic distribution of the test statistic  $W_n$  under the alternative hypotheses  $H_{1,n}$  and  $H_{1,n}^*$  are as given in (22) and (23) respectively.

**Theorem 11** *The asymptotic distribution of  $W_n$  is given by*

$$i) \ W_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi_r^2 \left( \mathbf{d}^T \mathbf{M}(\boldsymbol{\theta}_0) \left[ \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0) \right]^{-1} \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{d} \right) \text{ under } H_{1,n} \text{ given in (22).}$$

$$ii) \ W_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi_r^2 \left( \boldsymbol{\delta}^T \left[ \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0) \right]^{-1} \boldsymbol{\delta} \right) \text{ under } H_{1,n}^* \text{ given in (23).}$$

**Proof.** See the Appendix. ■

#### 4.1 Normal case

Under the  $N(\mu, \sigma^2)$  model, consider the problem of testing

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0, \quad (27)$$

where  $\sigma^2$  is an unknown nuisance parameter. In this case the parameter space is given by  $\Theta = \{(\mu, \sigma) \in \mathbb{R}^2 | \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ , and the parameter space under the null distribution is  $\Theta_0 = \{(\mu, \sigma) \in \mathbb{R}^2 | \mu = \mu_0, \sigma \in \mathbb{R}^+\}$ . If we consider the function  $m(\boldsymbol{\theta}) = \mu - \mu_0$ , where  $\boldsymbol{\theta} = (\mu, \sigma)^T$ , the null hypothesis  $H_0$  can be written as

$$H_0 : m(\boldsymbol{\theta}) = 0.$$

We observe that in our case  $\mathbf{M}(\boldsymbol{\theta}) = (1, 0)^T$ . Based on (4) and taking into account the fact that  $f_{\boldsymbol{\theta}}(x)$  is the normal density with mean  $\mu$  and variance  $\sigma^2$ , the estimator  $\hat{\boldsymbol{\theta}}_\beta = (\hat{\mu}_\beta, \hat{\sigma}_\beta)^T$  of  $\boldsymbol{\theta}$  is the solution of the system of nonlinear equations

$$\begin{cases} \frac{\partial}{\partial \mu} \frac{1}{\sigma^\beta (2\pi)^{\frac{\beta}{2}}} \left( \frac{1}{n\beta} \sum_{i=1}^n \exp \left\{ -\frac{\beta}{2} \left( \frac{X_i - \mu}{\sigma} \right)^2 \right\} - \frac{1}{(1+\beta)^{3/2}} \right) = 0, \\ \frac{\partial}{\partial \sigma} \frac{1}{\sigma^\beta (2\pi)^{\frac{\beta}{2}}} \left( \frac{1}{n\beta} \sum_{i=1}^n \exp \left\{ -\frac{\beta}{2} \left( \frac{X_i - \mu}{\sigma} \right)^2 \right\} - \frac{1}{(1+\beta)^{3/2}} \right) = 0. \end{cases}$$

Simple calculations yield the expressions

$$\mathbf{J}_\beta(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{1+\beta} (2\pi)^{\beta/2} \sigma^{2+\beta}} \begin{pmatrix} \frac{1}{1+\beta} & 0 \\ 0 & \frac{\beta^2+2}{(1+\beta)^2} \end{pmatrix},$$

and

$$\mathbf{K}_\beta(\boldsymbol{\theta}) = \frac{1}{\sigma^{2+2\beta} (2\pi)^\beta} \left( \frac{1}{(1+2\beta)^{3/2}} \begin{pmatrix} 1 & 0 \\ 0 & \frac{4\beta^2+2}{1+2\beta} \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & \frac{\beta^2}{(1+\beta)^3} \end{pmatrix} \right).$$

Therefore,

$$\mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0) = \sigma^2 \frac{(\beta+1)^3}{(2\beta+1)^{\frac{3}{2}}},$$

and

$$\begin{aligned} W_n &= nm^T(\hat{\boldsymbol{\theta}}_\beta) \left( \mathbf{M}^T(\hat{\boldsymbol{\theta}}_\beta) \mathbf{J}_\beta^{-1}(\hat{\boldsymbol{\theta}}_\beta) \mathbf{K}_\beta(\hat{\boldsymbol{\theta}}_\beta) \mathbf{J}_\beta^{-1}(\hat{\boldsymbol{\theta}}_\beta) \mathbf{M}(\hat{\boldsymbol{\theta}}_\beta) \right)^{-1} m(\hat{\boldsymbol{\theta}}_\beta) \\ &= n \frac{(\hat{\mu}_\beta - \mu_0)^2 (2\beta+1)^{\frac{3}{2}}}{\hat{\sigma}_\beta^2 (\beta+1)^3}, \end{aligned}$$

and on the basis of Theorem 9, we have

$$W_n = n \frac{(\hat{\mu}_\beta - \mu_0)^2 (2\beta+1)^{\frac{3}{2}}}{\hat{\sigma}_\beta^2 (\beta+1)^3} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi_1^2. \quad (28)$$

We observe that for  $\beta = 0$  we get the classical Wald test statistic for testing the hypothesis mentioned in (27).

## 4.2 Weibull case

While the normal model is the most important model where the test statistic described in Section 4.1 would be useful, it is also important to explore the applicability of these tests in other models to demonstrate the general nature of the method. In testing composite hypotheses, therefore, we will include numerical results based on the Weibull distribution in our subsequent numerical study, together with the results on the normal model. Here we describe the Wald-type test statistic for the Weibull case. The probability density function of  $\mathcal{W}(\sigma, p)$ , a two parameter Weibull distribution, is given by

$$f_{\boldsymbol{\theta}}(x) = \frac{p}{\sigma} \left(\frac{x}{\sigma}\right)^{p-1} \exp\left\{-\left(\frac{x}{\sigma}\right)^p\right\}, \quad x > 0,$$

where  $\boldsymbol{\theta} = (\sigma, p)^T$ , and the parameter space is given by  $\Theta = \{(\sigma, p) | \sigma \in \mathbb{R}^+, p \in \mathbb{R}^+\}$ . We are interested in testing

$$H_0 : \sigma = \sigma_0 \text{ against } H_1 : \sigma \neq \sigma_0, \quad (29)$$

where  $p$  is a nuisance parameter. Let us consider the function  $m(\boldsymbol{\theta}) = \sigma - \sigma_0$ . Then, as in the normal case which was considered in Section 4.1, the null hypothesis  $H_0$  can be written as

$$H_0 : m(\boldsymbol{\theta}) = 0,$$

and  $\mathbf{M}(\boldsymbol{\theta}) = (1, 0)^T$ .

Let us define

$$\xi_{\alpha, \beta}(\boldsymbol{\theta}) = \int_0^\infty \left(\frac{x}{\sigma}\right)^\alpha f_{\boldsymbol{\theta}}^\beta(x) dx,$$

and

$$\eta_{\alpha, \beta, \gamma}(\boldsymbol{\theta}) = \int_0^\infty \left(\frac{x}{\sigma}\right)^\alpha \left[\log\left(\frac{x}{\sigma}\right)\right]^\beta f_{\boldsymbol{\theta}}^\gamma(x) dx.$$

It can be shown that

$$\xi_{\alpha, \beta}(\boldsymbol{\theta}) = \left(\frac{p}{\sigma}\right)^{\beta-1} \beta^{-\frac{\beta p - \beta + \alpha + 1}{p}} \Gamma\left(\frac{\beta p - \beta + \alpha + 1}{p}\right), \quad (30)$$

and

$$\eta_{\alpha, \beta, \gamma}(\boldsymbol{\theta}) = \sigma \left(\frac{p}{\sigma}\right)^\gamma \int_0^\infty y^{\alpha + \gamma p - \gamma} (\log y)^\beta \exp(-\gamma y^p) dy, \quad (31)$$

where  $\Gamma(\cdot)$  denote the gamma function. Note that  $\xi_{\alpha, \gamma}(\boldsymbol{\theta}) = \eta_{\alpha, 0, \gamma}(\boldsymbol{\theta})$ . For  $\beta \neq 0$  the value of  $\eta_{\alpha, \beta, \gamma}(\boldsymbol{\theta})$  is calculated using numerical integration. Let us define

$$\mathbf{R}_\gamma(\boldsymbol{\theta}) = \int_0^\infty \mathbf{u}_{\boldsymbol{\theta}}(x) \mathbf{u}_{\boldsymbol{\theta}}^T(x) f_{\boldsymbol{\theta}}^\gamma(x) dx = \begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{21} \end{pmatrix},$$

where  $\mathbf{u}_{\boldsymbol{\theta}}(x)$ , the score function of the Weibull distribution, is given by

$$\mathbf{u}_{\boldsymbol{\theta}}(x) = \frac{\partial \log f_{\boldsymbol{\theta}}(x)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} -\frac{p}{\sigma} + \frac{p}{\sigma} \left(\frac{x}{\sigma}\right)^p \\ \frac{1}{p} + \log\left(\frac{x}{\sigma}\right) - \left(\frac{x}{\sigma}\right)^p \log\left(\frac{x}{\sigma}\right) \end{pmatrix}.$$

Then it can be shown that

$$r_{11} = \left(\frac{p}{\sigma}\right)^2 \{\xi_{0, \gamma}(\boldsymbol{\theta}) - 2\xi_{p, \gamma}(\boldsymbol{\theta}) + \xi_{2p, \gamma}(\boldsymbol{\theta})\},$$

$$r_{12} = \frac{p}{\sigma} \left\{ -\frac{1}{p} \xi_{0, \gamma}(\boldsymbol{\theta}) - \eta_{0, 1, \gamma}(\boldsymbol{\theta}) + 2\eta_{p, 1, \gamma}(\boldsymbol{\theta}) + \frac{1}{p} \xi_{p, \gamma}(\boldsymbol{\theta}) - \eta_{2p, 1, \gamma}(\boldsymbol{\theta}) \right\},$$

and

$$r_{22} = \frac{1}{p^2} \xi_{0, \gamma}(\boldsymbol{\theta}) + \eta_{0, 2, \gamma}(\boldsymbol{\theta}) + \eta_{2p, 2, \gamma}(\boldsymbol{\theta}) + \frac{2}{p} \eta_{0, 1, \gamma}(\boldsymbol{\theta}) - 2\eta_{p, 2, \gamma}(\boldsymbol{\theta}) - \frac{2}{p} \eta_{p, 1, \gamma}(\boldsymbol{\theta}).$$



Now

$$\mathbf{J}_\gamma(\boldsymbol{\theta}) = \int_0^\infty \mathbf{u}_\theta(x) \mathbf{u}_\theta^T(x) f_\theta^{1+\gamma}(x) dx = \mathbf{R}_{1+\gamma}(\boldsymbol{\theta}), \quad (32)$$

$$\mathbf{K}_\gamma(\boldsymbol{\theta}) = \int_0^\infty \mathbf{u}_\theta(x) \mathbf{u}_\theta^T(x) f_\theta^{1+2\gamma}(x) dx = \mathbf{R}_{1+2\gamma}(\boldsymbol{\theta}). \quad (33)$$

Then some little algebra shows that the Wald-type test statistic is given by

$$\begin{aligned} W_n &= n(\hat{\sigma}_\beta - \sigma_0)^2 \left[ \mathbf{M}^T(\boldsymbol{\theta}) \mathbf{J}_\beta^{-1}(\hat{\sigma}_\beta, \hat{p}_\beta) \mathbf{K}_\beta(\hat{\sigma}_\beta, \hat{p}_\beta) \mathbf{J}_\beta^{-1}(\hat{\sigma}_\beta, \hat{p}_\beta) \mathbf{M}(\boldsymbol{\theta}) \right]^{-1} \\ &= \frac{n(\hat{\sigma}_\beta - \sigma_0)^2 \left[ r_{\beta+1}^{(1,1)}(\hat{\sigma}_\beta, \hat{p}_\beta) r_{\beta+1}^{(2,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) - \left( r_{\beta+1}^{(1,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) \right)^2 \right]^2}{\begin{pmatrix} -r_{\beta+1}^{(2,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) & r_{\beta+1}^{(1,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) \end{pmatrix} \begin{pmatrix} r_{2\beta+1}^{(1,1)}(\hat{\sigma}_\beta, \hat{p}_\beta) & r_{2\beta+1}^{(1,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) \\ r_{2\beta+1}^{(1,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) & r_{2\beta+1}^{(2,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) \end{pmatrix} \begin{pmatrix} -r_{\beta+1}^{(2,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) \\ r_{\beta+1}^{(1,2)}(\hat{\sigma}_\beta, \hat{p}_\beta) \end{pmatrix}} \\ &= \frac{n \left\{ (\hat{\sigma}_\beta - \sigma_0) \left( \frac{\hat{p}_\beta}{\sigma_\beta} \right)^\beta \left[ \tilde{r}_{\beta+1}^{(1,1)}(\hat{p}_\beta) \tilde{r}_{\beta+1}^{(2,2)}(\hat{p}_\beta) - \left( \tilde{r}_{\beta+1}^{(1,2)}(\hat{p}_\beta) \right)^2 \right] \right\}^2}{\begin{pmatrix} -\tilde{r}_{\beta+1}^{(2,2)}(\hat{p}_\beta) & \tilde{r}_{\beta+1}^{(1,2)}(\hat{p}_\beta) \end{pmatrix} \begin{pmatrix} \tilde{r}_{2\beta+1}^{(1,1)}(\hat{p}_\beta) & \tilde{r}_{2\beta+1}^{(1,2)}(\hat{p}_\beta) \\ \tilde{r}_{2\beta+1}^{(1,2)}(\hat{p}_\beta) & \tilde{r}_{2\beta+1}^{(2,2)}(\hat{p}_\beta) \end{pmatrix} \begin{pmatrix} -\tilde{r}_{\beta+1}^{(2,2)}(\hat{p}_\beta) \\ \tilde{r}_{\beta+1}^{(1,2)}(\hat{p}_\beta) \end{pmatrix}}, \end{aligned}$$

where

$$\begin{aligned} \tilde{r}_\gamma^{(1,1)}(\hat{p}_\beta) &= \varepsilon_{0,\gamma}(\hat{p}_\beta) - 2\varepsilon_{\hat{p}_\beta,\gamma}(\hat{p}_\beta) + \varepsilon_{2\hat{p}_\beta,\gamma}(\hat{p}_\beta), \\ \tilde{r}_\gamma^{(1,2)}(\hat{p}_\beta) &= -\frac{1}{\hat{p}_\beta} \varepsilon_{0,\gamma}(\hat{p}_\beta) + \left( \log \hat{p}_\beta + \frac{1}{\hat{p}_\beta} \right) \varepsilon_{\hat{p}_\beta,\gamma}(\hat{p}_\beta) - \log \hat{p}_\beta \varepsilon_{2\hat{p}_\beta,\gamma}(\hat{p}_\beta) - \kappa_{0,1,\gamma}(\hat{p}_\beta) + \kappa_{\hat{p}_\beta,1,\gamma}(\hat{p}_\beta), \\ \tilde{r}_\gamma^{(2,2)}(\hat{p}_\beta) &= \frac{1}{\hat{p}_\beta} \varepsilon_{0,\gamma}(\hat{p}_\beta) - \frac{2}{\hat{p}_\beta} \log \hat{p}_\beta \varepsilon_{\hat{p}_\beta,\gamma}(\hat{p}_\beta) + (\log \hat{p}_\beta)^2 \varepsilon_{2\hat{p}_\beta,\gamma}(\hat{p}_\beta) \\ &\quad + \frac{2}{\hat{p}_\beta} \kappa_{0,1,\gamma}(\hat{p}_\beta) + \kappa_{0,2,\gamma}(\hat{p}_\beta) - 2 \log \hat{p}_\beta \kappa_{\hat{p}_\beta,1,\gamma}(\hat{p}_\beta), \\ \varepsilon_{\alpha,\gamma}(p) &= \gamma^{\frac{(p-1)\gamma+\alpha+1}{p}} \Gamma\left(\frac{(p-1)\gamma+\alpha+1}{p}\right), \\ \kappa_{\alpha,\delta,\gamma}(\hat{p}_\beta) &= \hat{p}_\beta \int_0^\infty (\log y)^\delta y^{(\hat{p}_\beta-1)\gamma+\alpha} \exp\{-\gamma y^{\hat{p}_\beta}\} dy. \end{aligned}$$

## 5 Simulation Study

In this section we present a simulation study to analyze the behavior of the Wald-type test statistics introduced in this paper with some classical procedures for the same problems. We pay special attention to the robustness problem.

### 5.1 Simple Null Hypothesis

We have proposed the Wald-type test statistic for the simple null hypothesis in (8). We will now study the performance of the test through simulation in case of the exponential model. The expression for the test statistic is simplified in (16). We want to test the hypothesis  $H_0 : \theta = 2$  against the alternative  $H_1 : \theta \neq 2$ . In the first case we have generated data from the  $\mathcal{Exp}(2)$  distribution, and the observed level (measured as the proportion of test statistics exceeding the chi-square critical value in 10,000 replications) are presented in Figure 1(a). We have taken three Wald-type test statistics for  $\beta = 0.1, 0.2$  and  $0.5$ , denoted by  $W(\beta)$ , and compared with the classical Wald test statistic. It is clear that the observed levels of all these tests are very close to the nominal level of  $0.05$ . Note that the classical Wald test statistic is a special case of the proposed family of test statistics corresponding to  $\beta = 0$ .

Next, the same hypotheses were tested when the data were generated from the  $\mathcal{Exp}(1)$  distribution. The observed power (obtained in a similar manner as above) of the test is presented in Figure 1(b) under the alternative hypothesis. It is noticed that the Wald test statistic performs best in terms of the power of the test, and for large  $\beta$  the test statistics show little bit poor performance. However, this discrepancy decreases rapidly with the sample size, and by the time  $n \geq 50$ , the observed powers are practically equal to one. In any case for  $\beta = 0.1$  or  $0.2$  the tests are almost as powerful as the classical Wald test.

To evaluate the stability of the level and the power of the tests under contamination, we repeated the simulations for  $H_0 : \theta = 2$  against  $H_0 : \theta \neq 2$  with data generated from the exponential mixture  $0.95\mathcal{Exp}(2) + 0.05\mathcal{Exp}(10)$ , and also tested the same hypotheses with data generated from  $0.95\mathcal{Exp}(1) + 0.05\mathcal{Exp}(10)$  distribution. The results are given in Figures 1(c) and 1(d) respectively.

In this case there is a huge inflation in the observed level of the classical Wald test statistic and somewhat smaller inflation in the tests with small values of  $\beta$ . But as  $\beta$  increases the resistant nature of the tests are clearly apparent. For larger values of  $\beta$  the levels turn out to be more acceptable. The opposite behavior is seen in case of power. There appears to be a complete breakdown in power for the classical Wald test, but the power remains much more stable for the other test statistics.

On the whole it appears to be fair to claim that for sample sizes equal to or larger than 40 the efficiency of many of our Wald-type tests are very close to the efficiency of the classical Wald test, at least in the example studied here, but the robustness properties of our tests are often significantly better than the classical Wald test in terms of maintaining the stability of both the level and power.

## 5.2 Composite Null Hypothesis

### 5.2.1 Normal Case

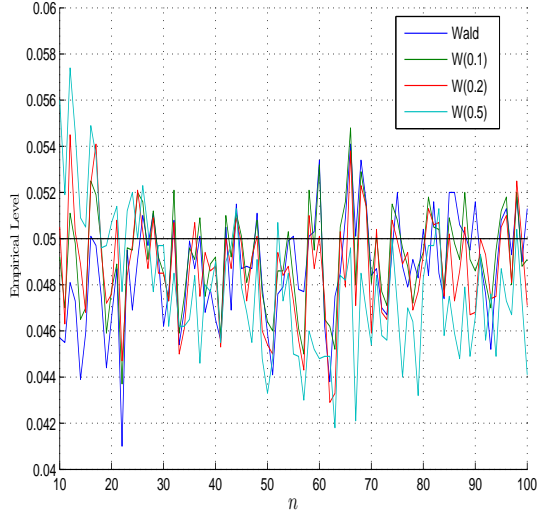
To explore the performance of our proposed test statistic in case of the composite null hypothesis, we have performed a simulation study for the family of test statistics given in Example 4.1. We considered the hypothesis  $H_0 : \mu = 0$  against the alternative  $H_1 : \mu \neq 0$  with  $\sigma^2$  unknown when data were generated from the  $\mathcal{N}(0, 1)$  distribution. Subsequently, the same hypotheses were tested when the data were generated from the  $\mathcal{N}(-1, 1)$  distribution. The results are given in Figures 2(a) and 2(b). In either case the nominal level was 0.05, and all tests were replicated 10,000 times.

It may be noticed that all the tests with large values of  $\beta$  are slightly liberal for very small sample sizes and lead to somewhat inflated observed levels. However, the observed levels settle down reasonably with increasing sample size. The observed powers of the tests as given in Figure 2(b) are, in fact, extremely close. In very small sample sizes our proposed test statistics have slightly higher power than the Wald test, but this must be a consequence of the observed levels of these tests being higher than the latter for such sample sizes. On the whole the proposed tests appear to be quite competitive to the ordinary Wald test for pure normal data.

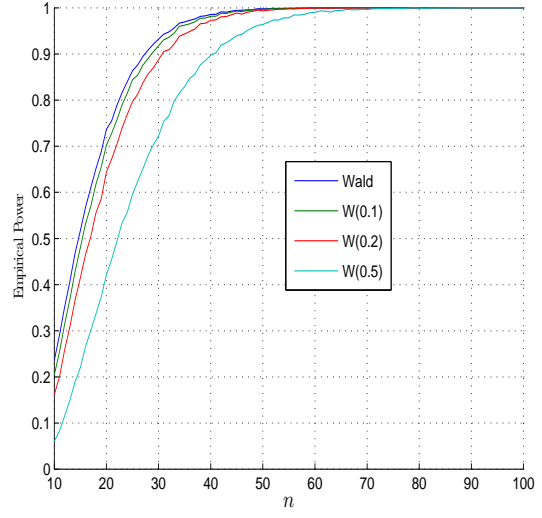
Now, we will show the performance of the proposed tests under contaminated data. So, we will test the same hypothesis, but the data have been generated from the  $0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(10, 1)$  distribution. The observed levels are shown in Figure 2(c). We notice that there is a drastic and severe inflation in the observed level of the ordinary Wald test. As  $\beta$  increases, however, the resistant nature of the tests are clearly apparent. By the time  $\beta = 0.2$ , the levels have already settled down to acceptable values.

Finally, we have generated data from the normal mixture  $0.9\mathcal{N}(-1, 1) + 0.1\mathcal{N}(10, 1)$ , and the power functions are plotted in Figure 2(d). There is a complete breakdown in power for the ordinary Wald test and the tests corresponding to the small values of  $\beta$ , but the power remains quite stable for values of  $\beta$  equal to 0.2 or greater.

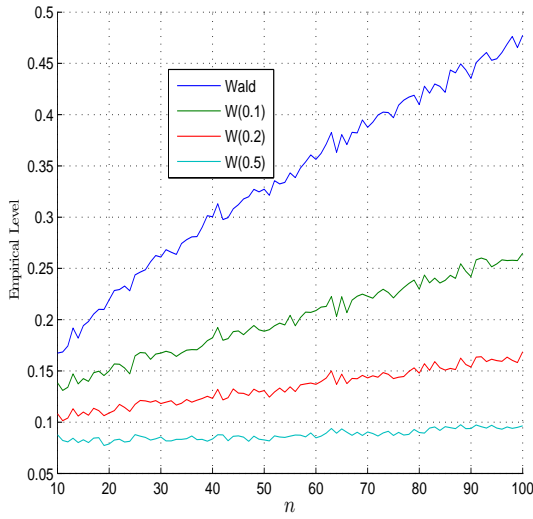
A similar conclusion as the previous simulation study can be drawn in this case. For sample sizes equal to or larger than 30 the efficiency of the Wald-type tests for  $\beta$  greater than 0.2 are very close to the efficiency of the ordinary Wald test, but the robustness properties of those tests are much better than the Wald test in terms of both the level and power.



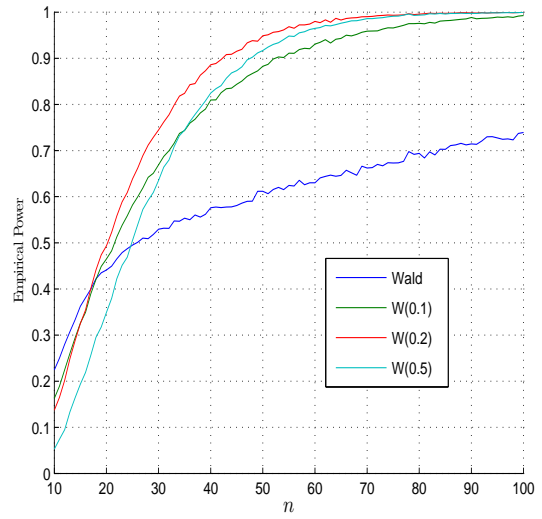
(a)



(b)

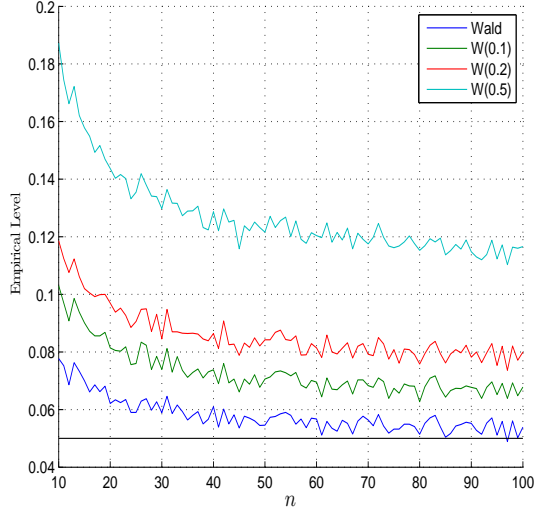


(c)

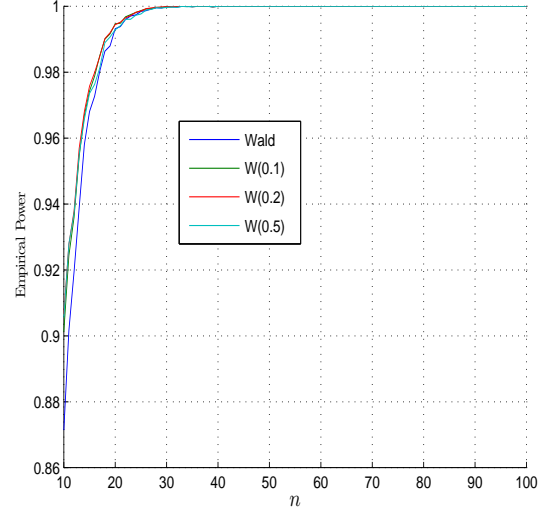


(d)

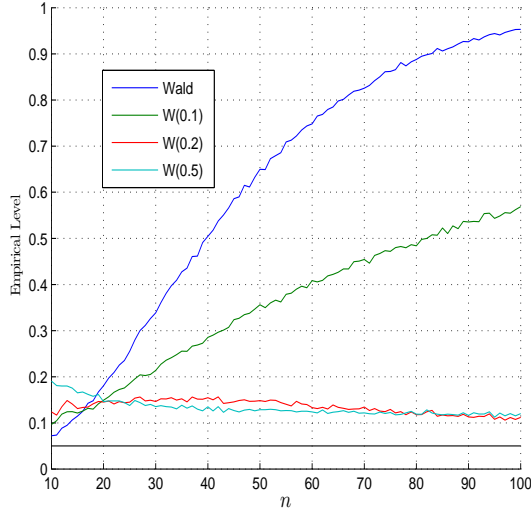
Figure 1: Simulated levels and powers of the Wald-type tests for pure and contaminated data in case of the exponential distribution.



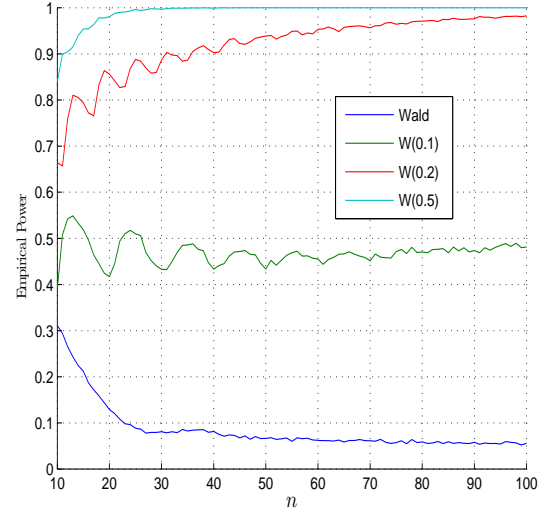
(a)



(b)



(c)



(d)

Figure 2: Simulated levels and powers of the Wald-type tests for pure and contaminated data in case of the normal distribution.

Table 1: Leukemia Data

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Count	23	7.5	43	26	60	105	100	170	54	70	94	320	350	1000	520	1000

### 5.2.2 Weibull Case

As we have mentioned before, it is important to demonstrate the properties of the proposed method in models other than the normal so that one has a better idea about the scope of the method. Accordingly we performed tests of composite hypotheses under the Weibull model in the spirit of Section 5.2.1. Let us consider the hypothesis defined in (29), where  $\sigma_0$  is taken to be 1.5. In the first study we have generated data from the  $\mathcal{W}(1.5, 1.5)$  distribution. The plot for the observed level for the hypothesis  $H_0 : \sigma = 1.5$  against the two sided alternative is given in Figure 3(a), where we have used 10,000 replications. Next the same hypotheses were tested when the data were generated from the  $\mathcal{W}(1, 1.5)$  distribution. The observed power function is plotted in Figure 3(b) for different values of  $\beta$ . The powers are remarkably close. In all cases the nominal level was 0.05.

To evaluate the stability of the level and the power of the tests under contamination, we repeated the tests with data generated from the Weibull mixture  $0.95\mathcal{W}(1.5, 1.5) + 0.05\mathcal{W}(10, 1.5)$ , and then from  $0.95\mathcal{N}(1, 1.5) + 0.05\mathcal{N}(10, 1.5)$ . In either case the first, larger component is our target. In Figure 3(c), the levels of the statistics under the contamination of first type are presented indicating the stability of levels for moderately large values of  $\beta$ . Figure 3(d) demonstrates the stability of powers under contaminated data of the second type for the same values of  $\beta$ .

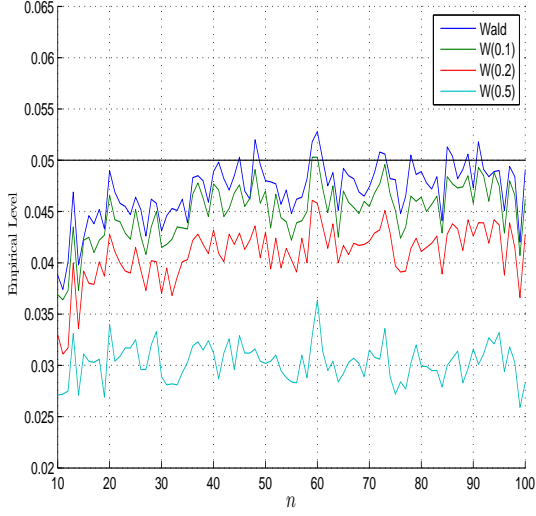
## 5.3 Real Data Examples

### 5.3.1 Leukemia Data

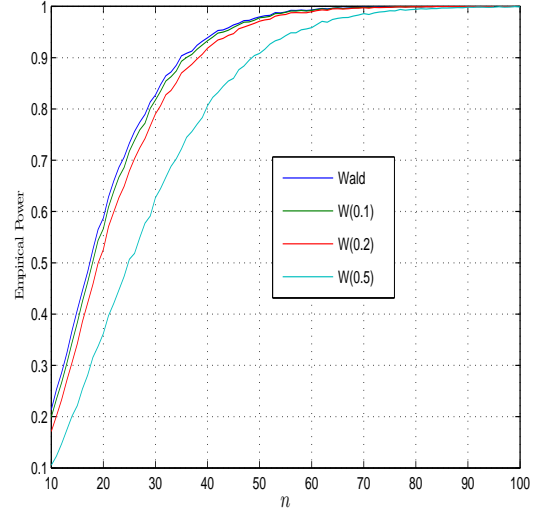
Let us consider the leukemia data set given in Tiku et al. (1986), and Gross and Clark (1975). Table 1 gives (100 times) the white blood cell counts of patients who had acute myelogenous leukemia. We assume an exponential distribution with parameter  $\theta$  models this data. One fact that is immediately noticeable is that there are two huge outlying observations at 1000 with respect to the exponential model, whereas other observations appear to be reasonable with respect to the same. The maximum likelihood estimator of  $\theta$  for the full data is 246.41, but if we delete the two outliers it comes out to be 138.75. We consider testing the null hypothesis  $H_0 : \theta = 140$  against  $H_1 : \theta \neq 140$  for the full data. The  $p$ -value of the classical Wald test is 0.0024, so in the presence of the outliers the ordinary Wald test rejects the null hypothesis. For the outlier deleted data the  $p$ -value of the ordinary Wald test becomes 0.9733. This extreme turnaround illustrates that the ordinary Wald test is highly influenced by very small proportion of outlying observations. On the other hand, the Wald-type tests for large values of  $\beta$  exhibit similar inference in all situations. Figure 4(a) shows the  $p$ -values of the Wald-type tests for the full data as well as for the outlier deleted data. This stable behavior of the test statistic based on the density power divergence for the full data approximately coincides with the stability of the minimum density power divergence estimates of  $\theta$  itself, obtained under an exponential model, which is presented in Figure 4(b). The minimum density power divergence estimators for the full data and the outlier deleted data are practically identical for  $\beta > 0.45$ . So the robustness of the test statistic is clearly linked to the robustness of the estimator.

### 5.3.2 Telephone-Fault Data

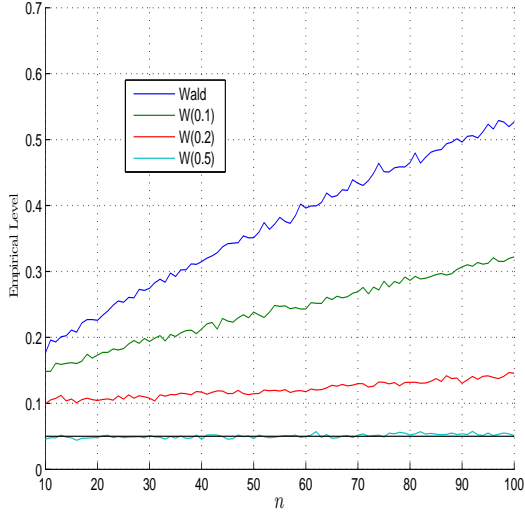
We consider the data on telephone line faults presented and analyzed by Welch (1987); the data were also analyzed by Simpson (1989). The data are given in Table 2, and consist of the ordered differences between the inverse test rates and the inverse control rates in 14 matched pairs of areas. A simple parametric approach would be to model these data as a random sample from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . One fact that is immediately noticeable is that the first observation



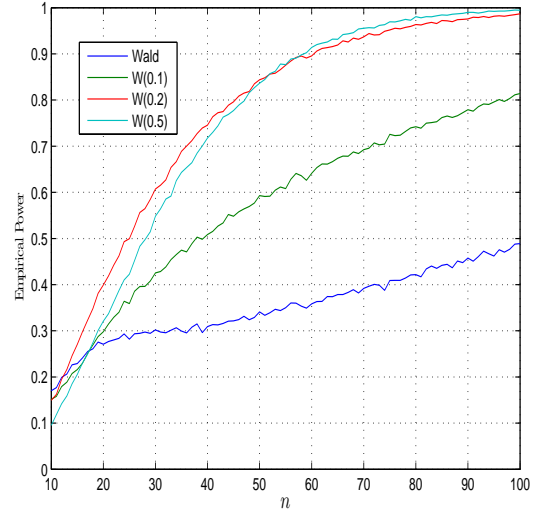
(a)



(b)



(c)



(d)

Figure 3: Simulated levels and powers of the Wald-type tests for pure and contaminated data in case of the Weibull distribution.

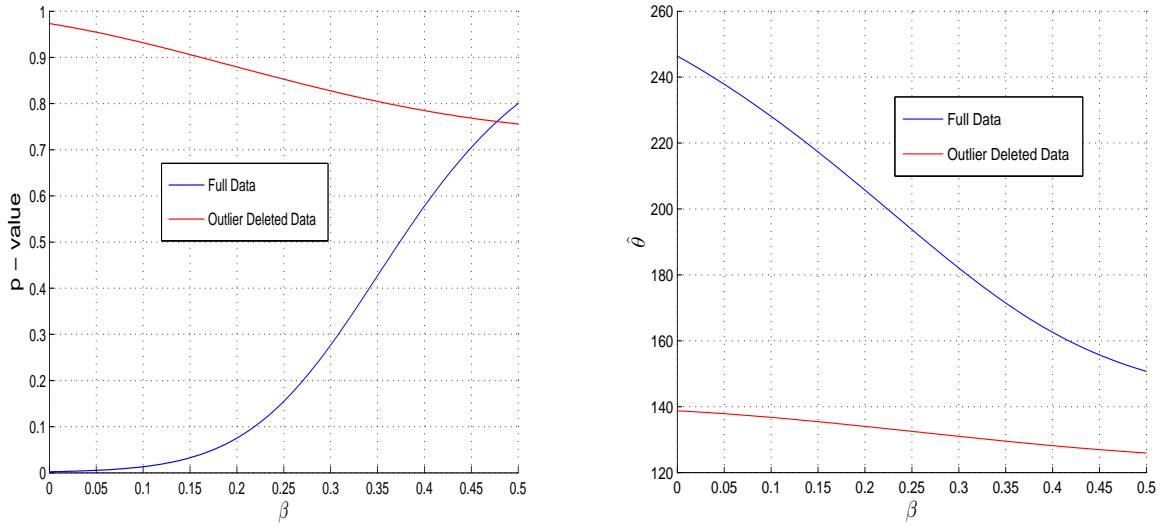


Figure 4: (a) Two sided  $p$ -values of the Wald-type tests, and (b) estimates of  $\theta$  for different values of  $\beta$  in case of the Leukemia data.

of this dataset is a huge outlier with respect to the normal model, while the remaining 13 observations appear to be reasonable with respect to the same. In Figure 5 we present a kernel density estimate for these data, the normal model fit based on the maximum likelihood estimates of  $\mu$  and  $\sigma$ , a normal model fit based on the minimum density power divergence estimates (with tuning parameter 0.15) of the normal parameters, and a normal model fit based on the minimum Hellinger distance estimates of the normal parameters (Simpson, 1989). The figure shows that if a small hump to the extreme left could be ignored, these data would have a nice unimodal structure which could be well modeled by an appropriate normal density. Apart from the minimum Hellinger distance estimates, such a normal density is provided in this figure by the minimum density power divergence estimates which correspond to  $\mu = 121.3$  and  $\sigma = 134.2$ . The maximum likelihood estimates, on the other hand, try to be inclusive and generate a result which neither models the outlier deleted data, nor provides a fit to the outlier component.

For the full data, the ordinary Wald test for the null hypothesis  $H_0 : \mu = 0$  against  $H_1 : \mu \neq 0$  fails to reject the null due to the presence of the large outlier (two sided  $p$ -value is 0.6584); however the robust Hellinger deviance test (Simpson, 1989) comfortably rejects the null (two sided  $p$ -value based on the chi-square null distribution is 0.0061), as does the ordinary Wald test based on the cleaned data after the removal of the large outlier (two sided  $p$ -value is 0.0076).

Table 2: Telephone-Fault Data

Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Difference	-988	-135	-78	3	59	83	93	110	189	197	204	229	289	310

Under the normal model, the usual estimates of  $\mu$  (and  $\sigma$ ) are highly inflated due to the presence of the large outlier, and as a result the likelihood ratio test under the normal model fails to reject the null hypothesis. From the robustness perspective, this is precisely what we will like to avoid, and here we demonstrate that proper choices of the tuning parameter within the class of tests developed in this paper achieve this goal. Here we analyze the performance of the Wald-type tests for different values of  $\beta$ . Figure 6(a) represents the  $p$ -values of the tests for different values of  $\beta$  in a region of interest. While it is clearly seen that the tests fail to reject the null hypothesis for these data at very small values of  $\beta$ , the decision turns around sharply, as  $\beta$  crosses and goes beyond 0.1. On the other hand, the

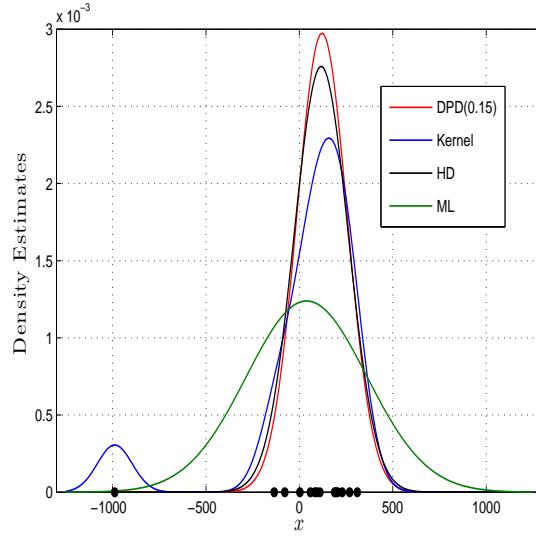


Figure 5: Kernel density estimate and different normal fits for the Telephone-Fault data.

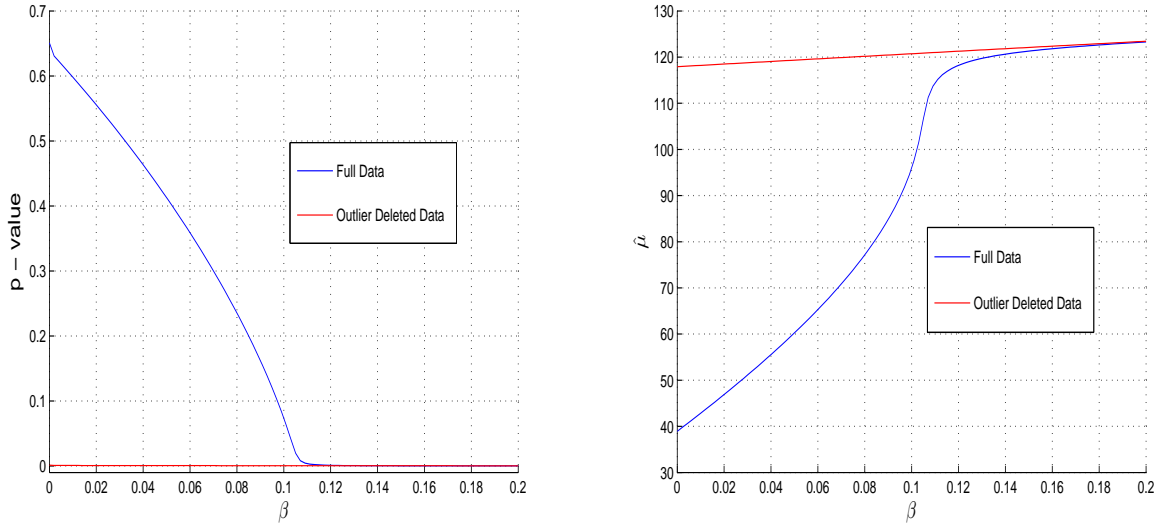


Figure 6: (a) Two sided  $p$ -values of the Wald-type tests, and (b) estimates of  $\mu$  for different values of  $\beta$  in case of the Telephone-Fault data.



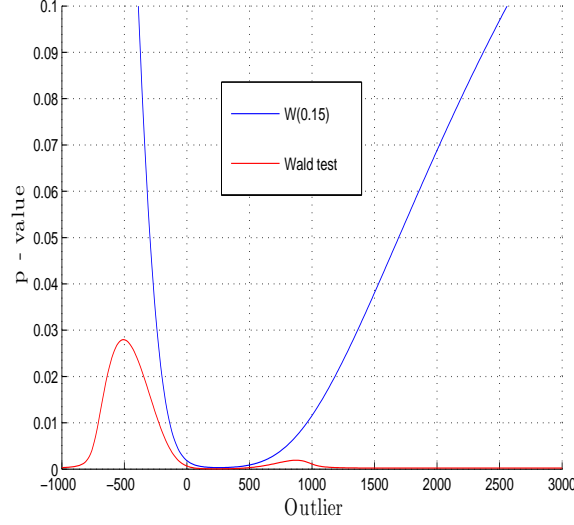


Figure 7: Two sided  $p$ -values for the tests for the mean for the Telephone-Fault data under the normal model against the first outlying observation.

$p$ -values of the same test based on the outlier deleted data remain stable, supporting rejection, at all values of  $\beta$  as seen in Figure 6(a). The stable behavior of the test statistic based on the density power divergence for the full data approximately coincides with the stability of the density power divergence estimate of  $\mu$  itself, obtained under a two-parameter normal model, which is presented in Figure 6(b). The minimum density power divergence estimators for the full data and the outlier deleted data are practically identical for  $\beta > 0.12$ . At least in this example, the robustness of the test statistic is clearly linked to the robustness of the estimator.

To further explore the robustness properties of the density power divergence tests we look at the two sided  $p$ -values for different values of the outlier. For this purpose we vary the first outlying observation in the range from  $-1000$  to  $3000$  by keeping the remaining 13 observations fixed. Figure 6(a) shows the corresponding  $p$ -values of the density power divergence tests with  $\beta = 0.15$  as well as the ordinary Wald test. It shows that initially the  $p$ -value of the density power divergence test with  $\beta = 0.15$  increases as the first observation moves away from the center of the data set, but after a certain limit the test gradually nullifies the effect of the outlier. On the other hand, the  $p$ -values of the Wald test keep on increasing with the outlier on either tail.

### 5.3.3 Darwin's Plant Fertilization Data

Charles Darwin had performed an experiment which may be used to determine whether self-fertilized plants and cross-fertilized plants have different growth rates. In this experiment pairs of *Zea mays* plants, one self and the other cross-fertilized, were planted in pots, and after a specific time period the height of each plant was measured. A particular sample of 15 such pairs of plants led to the paired differences (cross-fertilized minus self fertilized) presented in increasing order in Table 3 (see [Darwin, 1878](#)).

Table 3: Darwin's Plant Fertilization Data

Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Difference	-67	-48	6	8	14	16	23	24	28	29	41	49	56	60	75

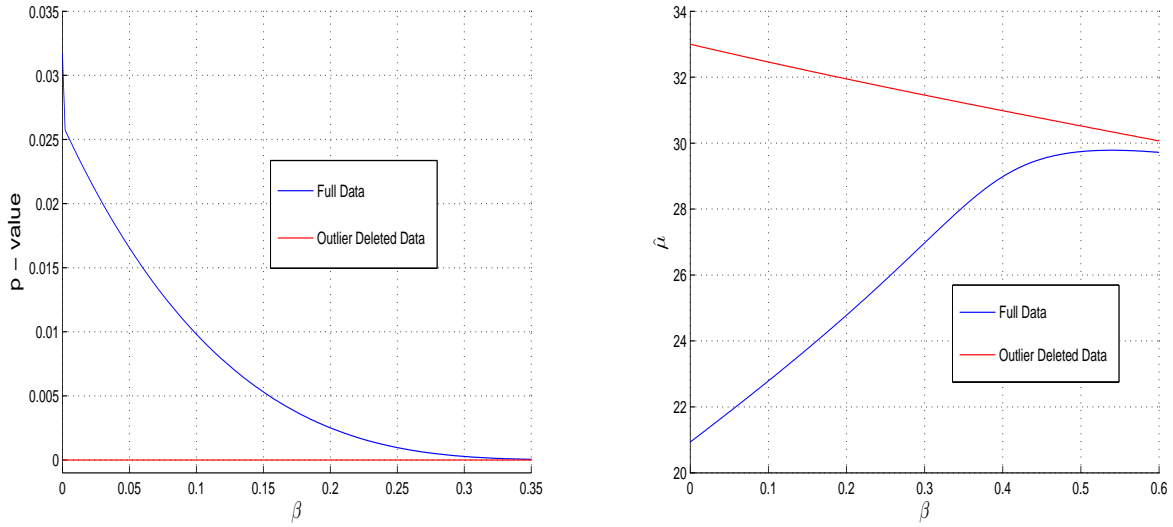


Figure 8: (a) Two sided  $p$ -values of the Wald-type tests, and (b) estimates of  $\mu$  for different values of  $\beta$  in case of Darwin's fertilization data.

As in the previous example, we assume a normal model for the paired differences and test  $H_0 : \mu = 0$  against  $H_1 : \mu \neq 0$ , i.e. we test whether the mean of the paired differences is different from zero. The unconstrained MDPDEs of  $\mu$  under the normal model corresponding to different values of the tuning parameter  $\beta$  are presented in Figure 8(b). The two negative paired differences appear to be geometrically well separated from the rest of the data, though they are perhaps not as huge outliers as the first observation in the Telephone-Fault data. These two observations do have a substantial impact on the parameter estimates and the test statistic for testing  $H_0$  using Wald-type tests with very small values of  $\beta$ , and it is instructive to compare the analysis to the case where these two outliers have been removed from the data. For small values of  $\beta$ , the two sided  $p$ -values of the test statistics are drastically different for the full data and outlier deleted cases (Figure 8(a)), but they get closer with increasing  $\beta$ , and they essentially coincide for  $\beta \geq 0.3$ . Once again this seems to be directly linked to the robustness of the parameter estimates; Figure 8(b), which also depicts the progression of the parameter estimates for the outlier deleted data, clearly demonstrates that.

### 5.3.4 One Sided Tests

In many parametric hypothesis testing problems including test for the mean under the normal model the case of real interest involves a one sided alternative. For the Telephone-Fault data problem, interest could lie in determining whether the mean fault rate is higher than zero (rather than simply whether it is different from zero). Darwin's interest in the fertilization problem was, presumably, to determine whether cross fertilization leads to a higher growth rate compared to self-fertilization; indeed the test performed on Darwin's data by R. A. Fisher (reported in [Fisher, 1966](#)) considers the one sided alternative. In this subsection we consider the relevant one sided alternatives for the two real data examples presented earlier in this section under the normal model. For this purpose we consider the signed Wald-type test statistic (the signed square root of the statistic presented in (28)) as in [Simpson \(1989\)](#), and determine the relevant (conservative) one sided  $p$ -value using the appropriate  $t$ -distribution. For a scalar parameter  $\theta$ , given the hypothesis  $H_0 : \theta = \theta_0$  against  $H_1 : \theta > \theta_0$ , the Wald-type statistic essentially turns out to be  $W_n = \frac{n^{1/2}(\hat{\theta}_\beta - \theta_0)}{\hat{\sigma}_\beta}$ , where  $\hat{\sigma}_\beta^2$  is the estimated variance of  $n^{1/2}\hat{\theta}_\beta$ . The test rejects at level  $\alpha$  when  $W_n$  exceeds the  $100(1 - \alpha)$ th quantile of the  $t$ -distribution with  $n - 1$  degrees of freedom (or the corresponding quantile of the standard normal distribution when  $n$  is large). Similarly, one gets a critical region in the lower tail for the less than type alternative.

**Telephone-Fault data:** [Simpson \(1989\)](#) has reported the one sided  $p$ -value (for the greater than alternative) in case of the Hellinger deviance test for the mean under the normal model to be 0.0085 for the full data and 0.0093 for the outlier deleted data. The one sided  $p$ -values for signed Wald-type test statistics corresponding to  $\beta = 0.15$  and  $\beta = 0.30$  are 0.0033 and 0.0020 respectively for the full data, and 0.0040 and 0.0030 respectively for the outlier deleted data, computed under the  $t(13)$  distribution. On the other hand, the  $p$ -values for the full data and the outlier deleted data are 0.33 and 0.004 for the ordinary one sided matched pairs Wald test. Clearly the outlier adversely affects the ordinary Wald test, but not the proposed robust Wald-type tests. The mean of the ordered differences between the inverse test rates and inverse control rates does appear to be actually greater than zero.

**Darwins’s plant fertilization data:** In this case we want to test whether the mean of the paired differences (cross-fertilized minus self-fertilized) is zero against the greater than alternative. The ordinary Wald test performed by [Fisher \(1966\)](#) gives a statistic of 2.15, with a one sided  $p$ -value of 0.025 computed under the  $t(14)$  distribution; the corresponding outlier deleted statistic has a one sided  $p$ -value of 0.00007. On the other hand, the one sided  $p$ -values for the signed Wald test statistics for  $\beta = 0.15$  and  $\beta = 0.30$  are 0.0145, 0.0027 for the full data, and 0.0001, 0.0001 for the outlier deleted data. Clearly the two large outliers are influencing the decision of the Wald test, or Wald-type tests for very small values of  $\beta$ , to the extent that the decision is reversed from what would have been obtained with the outlier cleaned data at 1% level of significance; however larger values of  $\beta$  lead to a more consistent behavior of the tests. This data set requires stronger downweighting compared to the Telephone-Fault data, as the outliers here are less extreme, and therefore more difficult to identify. Under a suitable robust test, it does appear that the mean growth of cross-fertilized plants is higher than that of the self-fertilized plants.

## 6 Concluding Remarks

In this paper we have constructed generalized Wald-type tests based on minimum density power divergence estimators proposed by [Basu et al. \(1998\)](#). These tests do not require any intermediate smoothing technique as in the case of the Hellinger deviance test, so our proposed techniques appear to stand out among robust tests of hypotheses based on minimum distance methods. Wald-type tests are easy to construct, and avoid the problem of determining the rejection region based on linear combinations of chi-square distributions as in the case of [Basu et al. \(2013, 2014\)](#). These class of tests have a huge scope of application. We have demonstrated improved results in different models including normal, exponential and Weibull distributions. Some real data are also analyzed, and we note that our methods can be useful in detecting either kind of incorrect decision caused by a small number of extreme outliers. This is exemplified by the analysis of the Leukemia data, where the classical test leads to a false positive, and the analysis of the Telephone-Fault data where the classical test fails to detect a true positive. We trust that the proposed tests will be very useful practical tools for statistical data analysis.

**Acknowledgments:** This work was partially supported by Grant MTM 2009-10072.

## References

- A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.
- A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*. CRC Press, Boca Raton, FL, 2011.
- A. Basu, A. Mandal, N. Martin, and L. Pardo. Testing statistical hypotheses based on the density power divergence. *Ann. Inst. Statist. Math.*, 65(2):319–348, 2013.
- A. Basu, A. Mandal, N. Martin, and L. Pardo. Density power divergence tests for composite null hypotheses. *arXiv preprint arXiv:1403.0330*, 2014.

- C. Darwin. *The Effects of Cross and Self Fertilization in the Vegetable Kingdom*. John Murray, London, 1878.
- D. De Angelis and G. A. Young. Smoothing the bootstrap. *Internat. Statist. Rev.*, 60(1):45–56, 1992.
- R. Fisher. *The Design of Experiments*. Hafner Press, New York, 1966.
- D. A. S. Fraser. Most powerful rank-type tests. *Ann. Math. Statist*, 28:1040–1043, 1957.
- A. J. Gross and V. Clark. *Survival Distributions: Reliability Applications in the Biomedical Sciences*, volume 11. Wiley, New York, 1975.
- M. C. Jones, N. L. Hjort, I. R. Harris, and A. Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001.
- L. Pardo. *Statistical inference based on divergence measures*. Chapman & Hall/CRC, Boca Raton, FL, 2006.
- R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980.
- D. G. Simpson. Hellinger deviance tests: efficiency, breakdown points, and examples. *J. Amer. Statist. Assoc.*, 84(405):107–113, 1989.
- M. L. Tiku, W. Y. Tan, and N. Balakrishnan. *Robust inference*. Marcel Dekker, New York, 1986.
- W. J. Welch. Rerandomizing the median in matched-pairs designs. *Biometrika*, 74(3):609–614, 1987.

## Appendix

**Proof of Theorem 3:** A first order Taylor expansion of  $l(\boldsymbol{\theta})$  at  $\widehat{\boldsymbol{\theta}}_{\beta}$  around  $\boldsymbol{\theta}^*$  gives

$$l(\widehat{\boldsymbol{\theta}}_{\beta}) - l(\boldsymbol{\theta}^*) = \left( \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}^*) + o_p(\|\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}^*\|).$$

Now the result follows easily, since the asymptotic distribution of  $n^{1/2} \left( l(\widehat{\boldsymbol{\theta}}_{\beta}) - l(\boldsymbol{\theta}^*) \right)$  coincides with the asymptotic distribution of  $n^{1/2} \left( \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}^*)$ .

**Proof of Theorem 5:** We have

$$n^{1/2} (\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}_0) = n^{1/2} (\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}_n) + n^{1/2} (\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = n^{1/2} (\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}_n) + \mathbf{d}.$$

Under  $H_{1,n}$  it follows that

$$n^{1/2} (\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}_n) \xrightarrow[n \rightarrow \infty]{L} N(\mathbf{0}, \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0)),$$

and

$$n^{1/2} (\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}_0) \xrightarrow[n \rightarrow \infty]{L} N(\mathbf{d}, \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0)).$$

On the other hand

$$\mathbf{W}_n = \mathbf{X}^T \mathbf{X},$$

where

$$\mathbf{X} = \left( \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \right)^{-1/2} n^{1/2} (\widehat{\boldsymbol{\theta}}_{\beta} - \boldsymbol{\theta}_0),$$

and under  $H_{1,n}$

$$\mathbf{X} \xrightarrow[n \rightarrow \infty]{L} N \left( \left( \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \right)^{-1/2} \mathbf{d}, \mathbf{I}_{p \times p} \right).$$

Here  $\mathbf{I}_{p \times p}$  is the identity matrix of order  $p$ . Therefore

$$W_n = \mathbf{X}^T \mathbf{X} \xrightarrow[n \rightarrow \infty]{L} \chi_p^2(\delta)$$

with  $\delta = \mathbf{d}^T \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{d}$ ; here  $\chi_p^2(\delta)$  denotes a non-central chi-square distribution with  $p$  degrees of freedom and non-centrality parameter  $\delta$ .

**Proof of Theorem 9:** Let  $\boldsymbol{\theta}_0 \in \Theta_0$  be the true value of  $\boldsymbol{\theta}$ . Using a Taylor series expansion we get

$$\begin{aligned} \mathbf{m}(\hat{\boldsymbol{\theta}}_\beta) &= \mathbf{m}(\boldsymbol{\theta}_0) + \mathbf{M}(\boldsymbol{\theta}_0)^T (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_0\|) \\ &= \mathbf{M}(\boldsymbol{\theta}_0)^T (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_0\|), \end{aligned} \quad (34)$$

because from equation (17) we have  $\mathbf{m}(\boldsymbol{\theta}_0) = \mathbf{0}_r$ . Now, under  $H_0$ ,

$$n^{1/2} (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_0) \xrightarrow[n \rightarrow \infty]{L} N_p(\mathbf{0}_p, \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0)).$$

Therefore, from equation (34) we get, under  $H_0$ ,

$$n^{1/2} \mathbf{m}(\hat{\boldsymbol{\theta}}_\beta) \xrightarrow[n \rightarrow \infty]{L} N_r(\mathbf{0}_r, \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0)).$$

As  $\text{Rank}(\mathbf{M}(\boldsymbol{\theta})) = r$ , we get

$$n \mathbf{m}^T(\hat{\boldsymbol{\theta}}_\beta) \left[ \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0) \right]^{-1} \mathbf{m}(\hat{\boldsymbol{\theta}}_\beta) \xrightarrow[n \rightarrow \infty]{L} \chi_r^2.$$

Now  $\mathbf{M}^T(\hat{\boldsymbol{\theta}}_\beta) \mathbf{J}_\beta^{-1}(\hat{\boldsymbol{\theta}}_\beta) \mathbf{K}_\beta(\hat{\boldsymbol{\theta}}_\beta) \mathbf{J}_\beta^{-1}(\hat{\boldsymbol{\theta}}_\beta) \mathbf{M}(\hat{\boldsymbol{\theta}}_\beta)$  is a consistent estimator of  $\mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0)$ . Hence, under  $H_0$ ,

$$n \mathbf{m}^T(\hat{\boldsymbol{\theta}}_\beta) \left[ \mathbf{M}^T(\hat{\boldsymbol{\theta}}_\beta) \mathbf{J}_\beta^{-1}(\hat{\boldsymbol{\theta}}_\beta) \mathbf{K}_\beta(\hat{\boldsymbol{\theta}}_\beta) \mathbf{J}_\beta^{-1}(\hat{\boldsymbol{\theta}}_\beta) \mathbf{M}(\hat{\boldsymbol{\theta}}_\beta) \right]^{-1} \mathbf{m}(\hat{\boldsymbol{\theta}}_\beta) \xrightarrow[n \rightarrow \infty]{L} \chi_r^2.$$

**Proof of Theorem 10:** We note that  $l^*(\hat{\boldsymbol{\theta}}_\beta, \hat{\boldsymbol{\theta}}_\beta)$  and  $l^*(\hat{\boldsymbol{\theta}}_\beta, \boldsymbol{\theta}^*)$  have same asymptotic distribution, because  $\hat{\boldsymbol{\theta}}_\beta \xrightarrow[n \rightarrow \infty]{p} \boldsymbol{\theta}^*$ . Now a first order Taylor expansion of  $l^*(\hat{\boldsymbol{\theta}}_\beta, \boldsymbol{\theta}^*)$  at  $\hat{\boldsymbol{\theta}}_\beta$  around  $\boldsymbol{\theta}^*$  gives

$$l^*(\hat{\boldsymbol{\theta}}_\beta, \boldsymbol{\theta}^*) - l^*(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \left( \frac{\partial l^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}^T (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}^*) + o_p(\|\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}^*\|).$$

So the theorem is proved from the following result

$$n^{1/2} (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}^*) \xrightarrow[n \rightarrow \infty]{L} N_p(\mathbf{0}_p, \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}^*) \mathbf{K}_\beta(\boldsymbol{\theta}^*) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}^*)).$$

**Proof of Theorem 11:** A Taylor series expansion of  $\mathbf{m}(\hat{\boldsymbol{\theta}}_\beta)$  around  $\boldsymbol{\theta}_n$  yields

$$\mathbf{m}(\hat{\boldsymbol{\theta}}_\beta) = \mathbf{m}(\boldsymbol{\theta}_n) + \mathbf{M}^T(\boldsymbol{\theta}_n) (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_n) + o(\|\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_n\|).$$

From (25) we have

$$\mathbf{m}(\hat{\boldsymbol{\theta}}_\beta) = n^{-1/2} \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{d} + \mathbf{M}^T(\boldsymbol{\theta}_n) (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_n) + o(\|\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_n\|) + o(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|). \quad (35)$$

Under  $H_{1,n}$  we get  $n^{1/2} (o(\|\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_n\|) + o(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|)) = o_p(1)$  and

$$n^{1/2} (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_n) \xrightarrow[n \rightarrow \infty]{L} N_p(\mathbf{0}_p, \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_\beta(\boldsymbol{\theta}_0) \mathbf{J}_\beta^{-1}(\boldsymbol{\theta}_0)).$$

So from (35) we have

$$n^{1/2} \mathbf{m}(\hat{\boldsymbol{\theta}}_{\beta}) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N_r(\mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{d}, \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0)).$$

From (26) we get, under  $H_{1,n}^*$ ,

$$n^{1/2} \mathbf{m}(\hat{\boldsymbol{\theta}}_{\beta}) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N_r(\boldsymbol{\delta}, \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0)).$$

We apply the following result concerning quadratic forms: If  $\mathbf{Z} \in N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ,  $\boldsymbol{\Sigma}$  is a symmetric projection of rank  $k$  and  $\boldsymbol{\Sigma} \boldsymbol{\mu} = \boldsymbol{\mu}$ , then  $\mathbf{Z}^T \mathbf{Z}$  is a chi-square distribution with  $k$  degrees of freedom and non-centrality parameter  $\boldsymbol{\mu}^T \boldsymbol{\mu}$ . Here the quadratic form is

$$W_n = \mathbf{Z}^T \mathbf{Z},$$

where

$$\mathbf{Z} = n^{1/2} \mathbf{m}(\hat{\boldsymbol{\theta}}_{\beta}) \left[ \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0) \right]^{-1/2}.$$

We know

$$\mathbf{Z} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N_r \left( \left[ \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0) \right]^{-1/2} \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{d}, \mathbf{I}_r \right),$$

where  $\mathbf{I}_r$  is the identity matrix of order  $r$ . Hence the application of the result is immediate. The non-centrality parameter is

$$\mathbf{d}^T \mathbf{M}(\boldsymbol{\theta}_0) \left[ \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{K}_{\beta}(\boldsymbol{\theta}_0) \mathbf{J}_{\beta}^{-1}(\boldsymbol{\theta}_0) \mathbf{M}(\boldsymbol{\theta}_0) \right]^{-1} \mathbf{M}^T(\boldsymbol{\theta}_0) \mathbf{d}.$$